High Quality Search Results for HEC

Aleatha Parker-Wood, Darrell D.E. Long

University of California, Santa Cruz

CRIS
NSF Center for Research in Intelligent Storage

ssrc
Storage Systems Research Center

CITRIS
Center for Information Technology Research in the Interest of Society

Baskin Engineering
UC Santa Cruz

Exascale indexing

- Exascale means billions to trillions of files, within an order of magnitude of the current web
- Much like the web in 1999 did, current efforts in indexing focus on a SQL-like query language (SPARQL, QUASAR...)
- Query capabilities are lagging behind scale

But I like SQL!

- Great for power users!
- Not so great for scientists
- Formal query languages are hard for novice users
- Queries are either under or over-specific
- Exploratory queries versus seeking queries
- Too much of a billion files is WAY too much

Empower Users

- Allow users to be only as specific as they need to be
- Degrade gracefully from formal queries to keyword search
- Give users the ability to explore as well as find
- Give users ranked search

Why can't we just use Google?

- Google relies on the innate structure of the web
- Links between pages are implicit endorsements
- Current file system structures:
  - Directories?
  - Hard/soft links?
  - File names?
  - Access times?
- None of these are actually endorsements
- Without this, Google (and other modern search engines) are <u>just another similarity search</u>

Better endorsements

- Really want a measure of how popular a file is
  - How often is it accessed?
  - How recently?
  - By how many users?
- Usage patterns are a clear endorsement
- We need new metadata to measure this

One simple implementation

- Probabilistic access counters
- Decayed over time to favor recently popular files
- Require only one new int as a metadata field
- Can be implemented very efficiently
- Simple, not very powerful
- Is different from most existing ranking algorithms
- Not very customizable
- Better than nothing

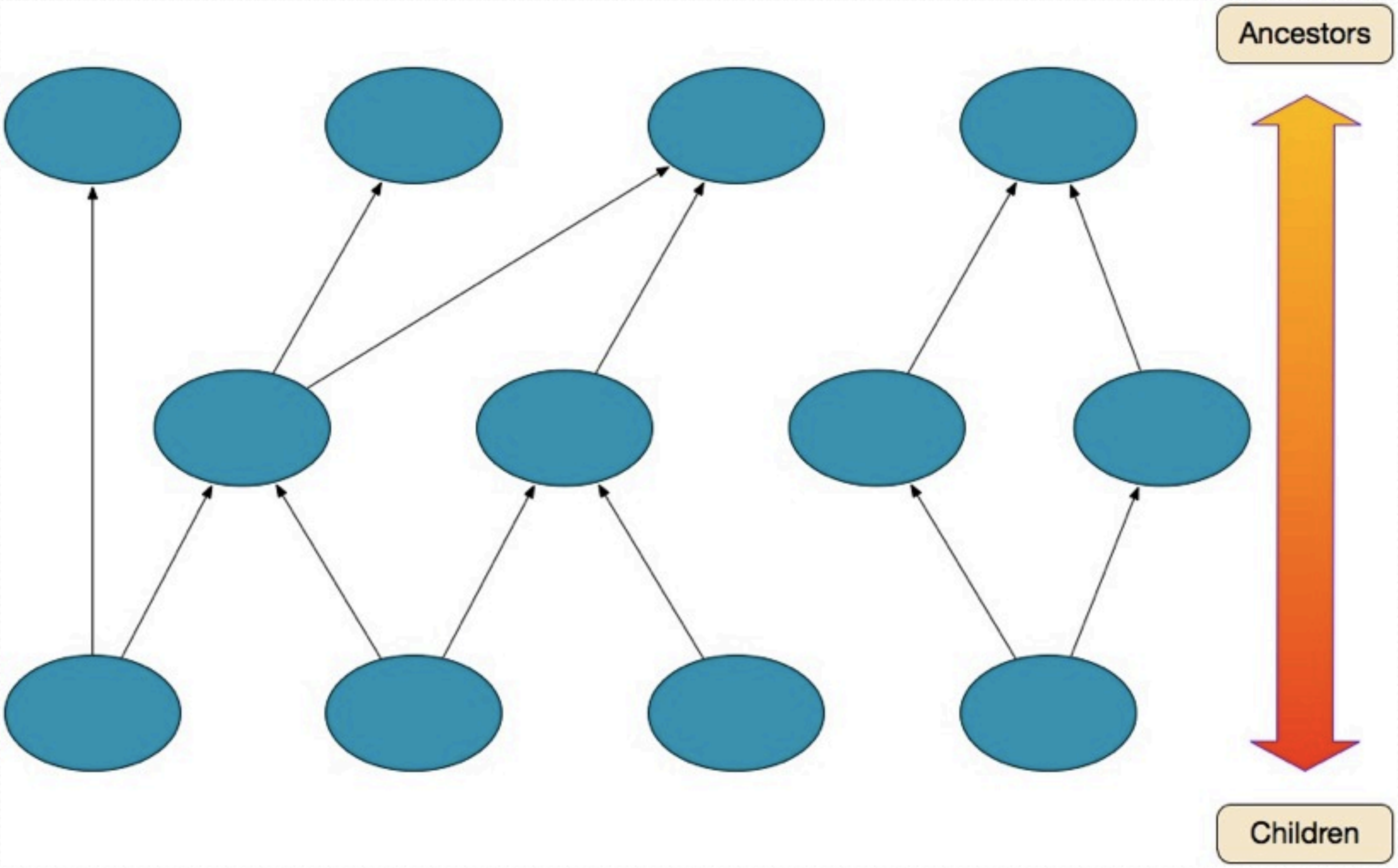
File system level provenance

- Not a new idea (Seltzer 2006, Cao 2005)
- Track data flow, file opens, files closes
- In combination with other metadata, can tell us:
  - How many people used a file
  - How recently
  - How often
- By interpreting provenance as endorsement, we can **leverage** existing ranking algorithms, and **create** new ones

P-rank

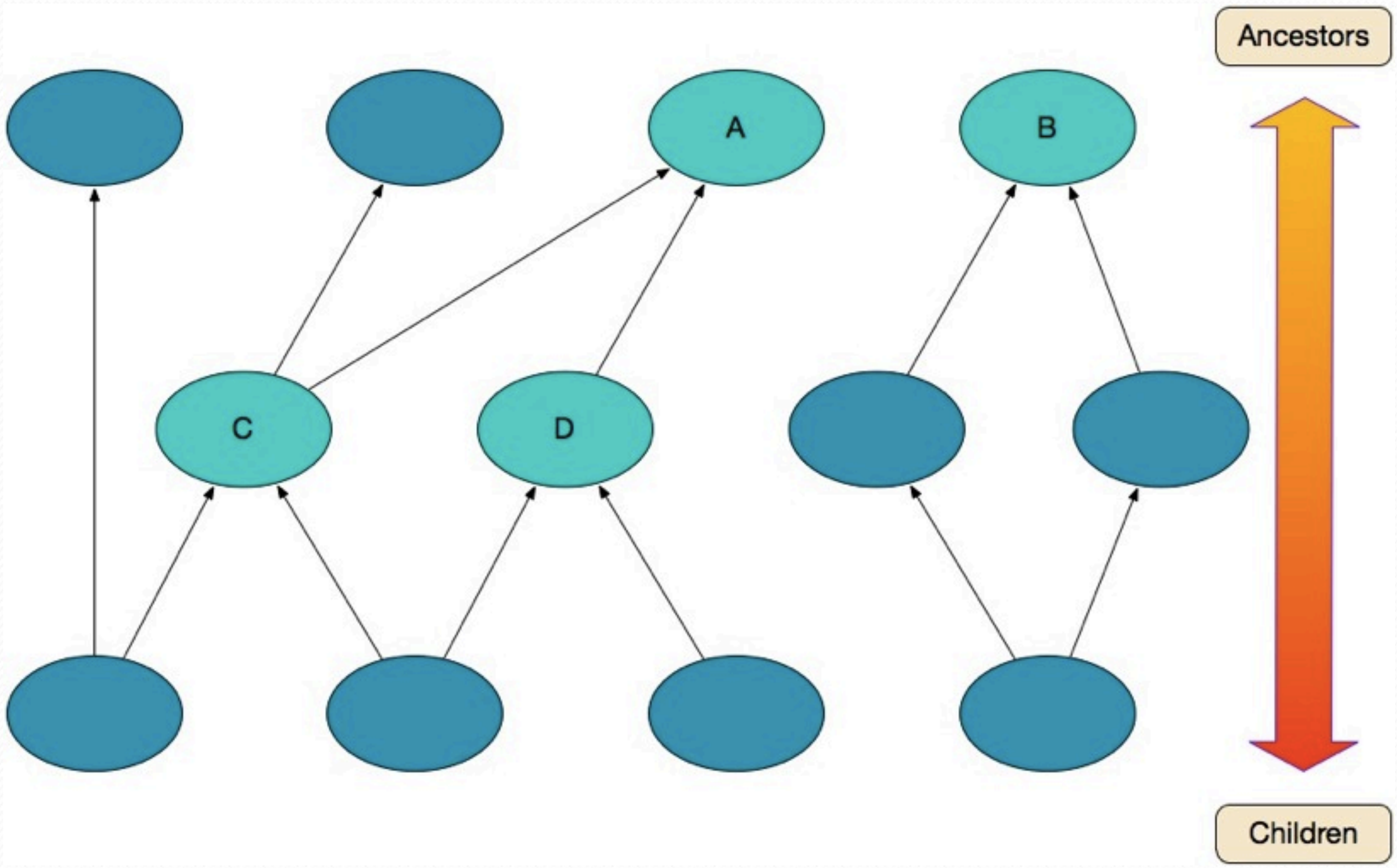
- Like Google’s PageRank, relies on matrix manipulation
- Modern HPC is very good at parallel matrix math
- Provenance is static once recorded
- Large parts of the matrix don't need recalculation

P-Rank



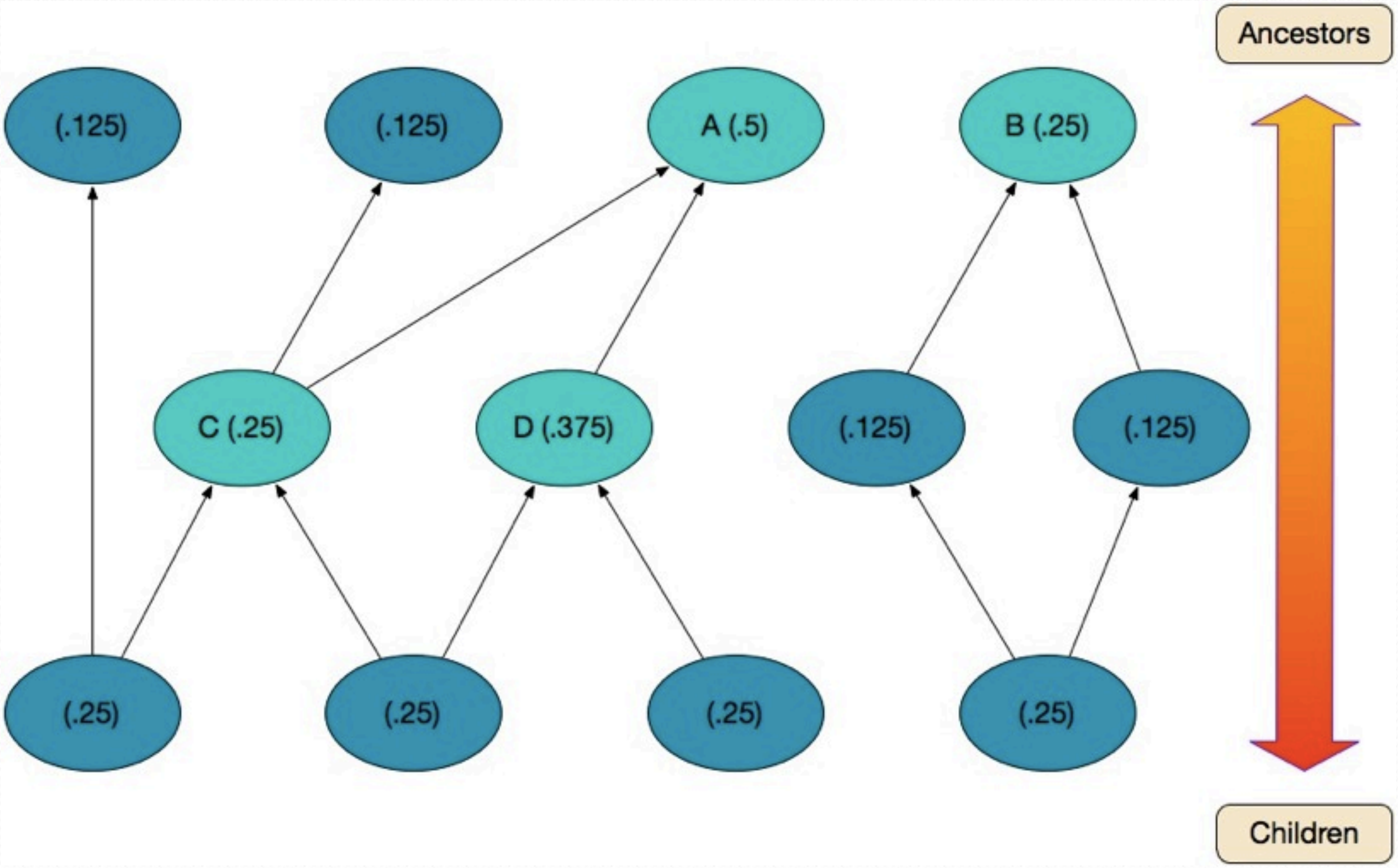
Example prov graph

P-Rank



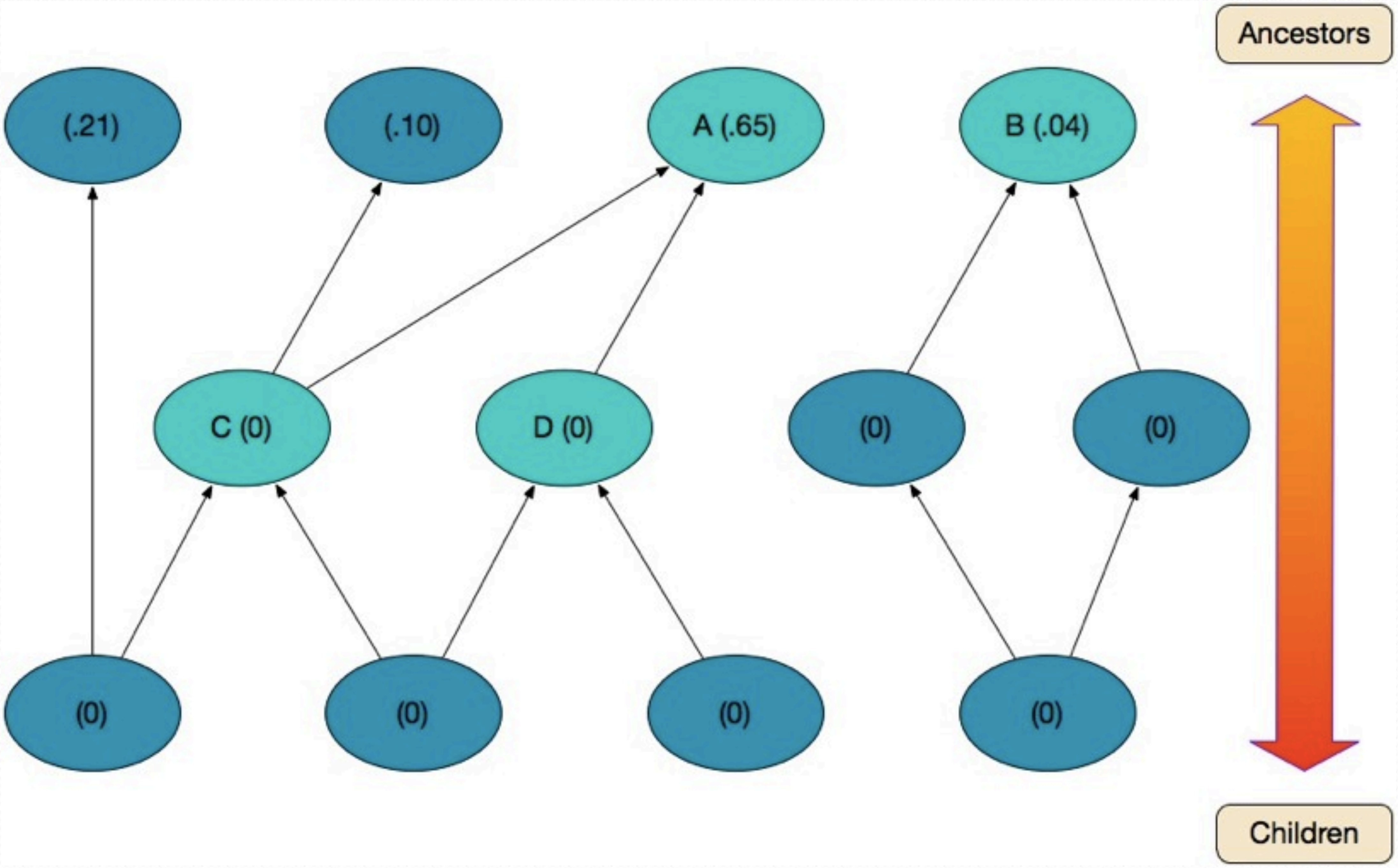
Call out popular nodes (high in-degree)

P-Rank



Add some probabilities (start at leaves, uniform transition along each link)

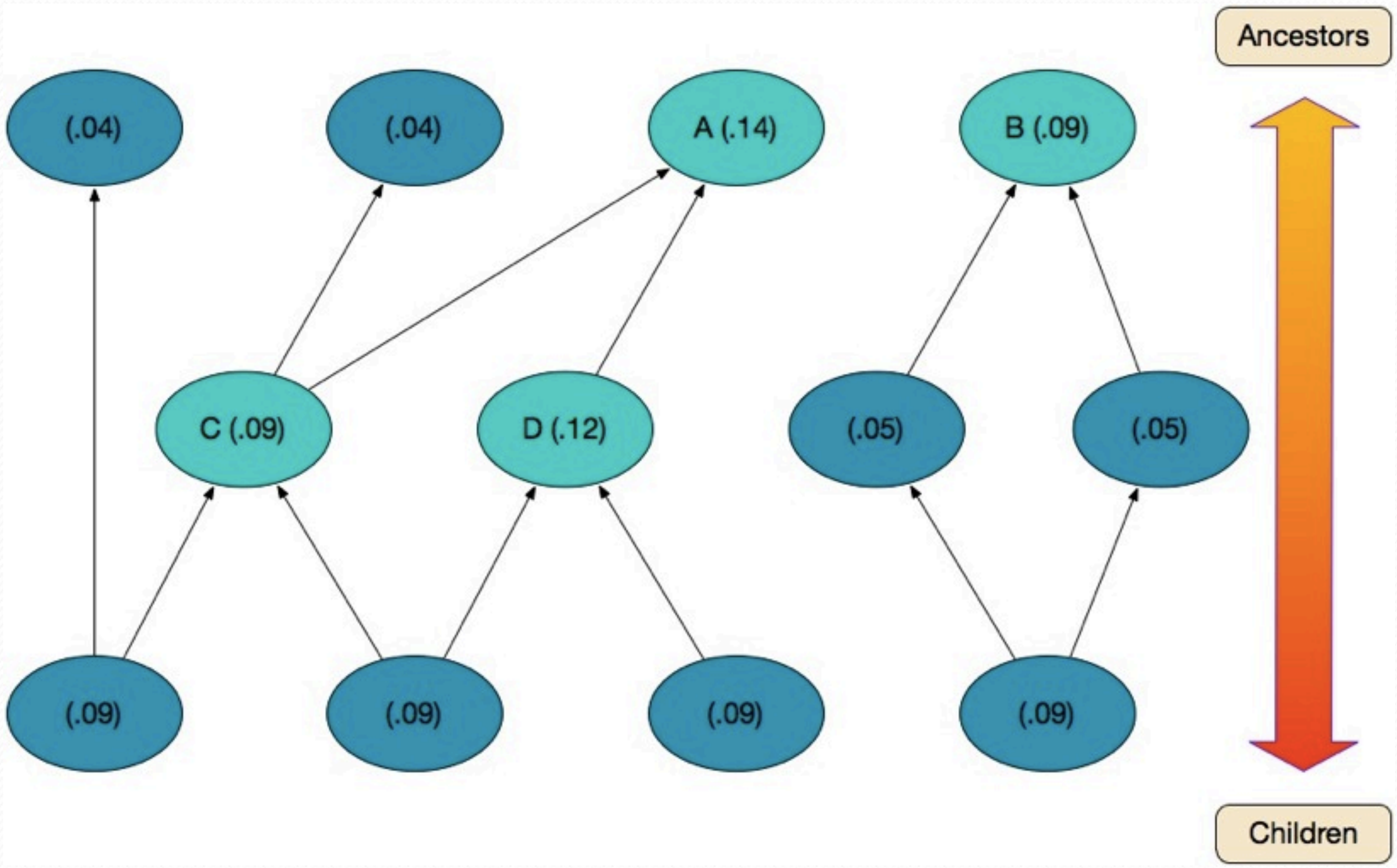
P-Rank



Monday, August 8, 2011

Pure page rank with a 25% teleport probability

P-Rank



Monday, August 8, 2011

Weighted page rank with leaf teleportation only

Personalization opportunities

- We have more information than Google does
- We can associate the *querier* with the *content creator*
- Having a rich graph allows us to do smarter queries

New types of queries

- Social network analysis
  - "Show me only my files"
  - “Find my working groups and boost files by them”
  - “Show me publicly visible files for my new team”
- "Show me codes compiled with this buggy library"
- Emigrant data forensics (Strong 2011)

Conclusion

- Ranked search has proven to be a powerful enabler on the web
- Has non-obvious performance benefits
- File systems lack structure for effective ranking
- With a modicum of metadata, we can do more
  - Ranked search
  - Personalized search
  - Powerful new queries